Statistics Questionnaire and answers

Q1 to Q9 have only one correct answer. Choose the correct option to answer your question.

1. Bernoulli random variables take (only) the values 1 and 0.
a) True

2. Which of the following theorem states that the distribution of averages of iid variables, properly normalized, becomes that of a standard normal as the sample size increases?
b) Central Mean Theorem

3. Which of the following is incorrect with respect to use of Poisson distribution?
d) All of the mentioned

4. Point out the correct statement.
d) All of the mentioned

5. _____ random variables are used to model rates.
c) Poisson

6. 10. Usually replacing the standard error by its estimated value does change the CLT.
b) False

7. 1. Which of the following testing is concerned with making decisions using data?
b) Hypothesis

8. 4. Normalized data are centered at_____and have units equal to standard deviations of the original data.
a) 0

9. Which of the following statement is incorrect with respect to outliers?
c) Outliers cannot conform to the regression relationship

Q10and Q15 are subjective answer type questions, Answer them in your own words briefly.

10. What do you understand by the term Normal Distribution?

Ans. Normal Distribution is continuous distribution in nature. Data like Salary, height, age can be taken as example for being continuous in nature and are normal distribution. In normal distribution every event is independent of each other ex: salary and height in a data set are independent of each other. In normal distribution Mean, Median and Mode are lined up such that Mean forms the centre of distribution, in normal distribution Mean=Median=Mode, because of this half of the results fall on either side of Mean and it is identifiable by bell shape. It is some times referred to as bell curve. Normal distribution is also called as Gaussian Distribution.

11. How do you handle missing data? What imputation techniques do you recommend?

Ans. Missing data can be handled by replacing the missing data with data from imputation techniques. Mean, Median, mode and constant value can be used as imputation techniques for replacing the missing data. In pandas for replacing the missing values fillna() method can be used with mean, median, mode and constant value . sklearn.impute.SimpleImputer can also be used for missing values imputation.

Box plots and distribution plots can be used for deciding the desired technique to be used for imputation among mean, median and mode. In limited data type of imputation can also be decided by observation, for example: If a model is to be built for chances of admission to a university, where the features provided are, GRE marks, TOEFL marks, CGPA, SOP etc. in case of GRE marks and TOEFL marks considering the mean imputation for handling missing data is apt as median and mode will not provide proper results for building model.

12. What is A/B testing?

Ans. A/B testing is basically statistical hypothesis testing. It is an analytical method for making decisions that estimates visitors of website based on sample statistics. It is also known as split testing. It allows decision makers to choose best design by testing with two or more versions/ alternatives A and B of variable (web site, web page, page element etc) and show to different segments visitors. Version that increases business metrics is considered.

13. Is mean imputation of missing data acceptable practice?

Ans. Mean Imputation of missing data depends on the type of data on which imputation is carried, following cases are examples

i) *Data for salaries:* In this case, If missing salary an intern employee is replaced with mean imputation of salary, then the graph might be skewed and salary may be relatively high for the Intern. Hence mean imputation is not acceptable in this case.

ii) *Data on age and fitness:* In this case, If missing fitness of a four year old kid is replaced with mean imputation of fitness, then the fitness replaced may be relatively high for the Kid. Hence mean imputation is not acceptable in this case.

iii) As explained in question 11, if a model is built for admission probability of a university with predefined data and cap of minimum marks is provided, A mean imputation can be considered as acceptable practise in cases where GRE score is missing.

14. What is linear regression in statistics?

Ans. Linear Regression in statistics define the relationships between one explanatory variable and one outcome variable. Explanatory variables are features and are independent, outcome variables are labels and dependent variables. There are two types of regression based on variables

i) Simple Linear Regression when we have one independent variable and

ii) Multiple Regression when we have more independent variables.

15. What are the various branches of statistics?

Ans. There are two branches of Statistics i) Descriptive Statistics ii) Inferential Statistics

i) Descriptive Statistics: As the name suggests, if the data is describable then the statistics applied is Descriptive Statistics, examples of descriptive statistics are Average height, marks, quality of product etc

ii) Inferential Statistics: Inferential statistics is used where data cannot be specifically defined.
For example, If in a state elections, exit polls data is to be obtained by statistics, it is impossible to get all data or assess all data, In such cases random samples of population is considered and statistics is applied on those random samples. (Here population refers to the type of data on which statistics is done, in case of election votes are considered population)